VE401 Recitation Class Note12

Simple Linear Regression

Chen Siyi
siyi.chen_chicy@sjtu.edu.cn

1 Basic Model

Setting and Assumptions:

(i) A dependent variable Y , assume to follow a normal distribution.

(ii) An independent variable X, which we can assume to either be a non-random parameter or a random variable measured precisely, without any error or uncertainty.

We want to describe $Y|x$

1.1 Simple Linear Regression Model

We assume that the mean $\mu_{Y|x}$ is given by

$$\mu_{Y|x} = \beta_0 + \beta_1 x \quad \text{for some } \beta_0, \beta_1 \in \mathbb{R}$$

This is called a simple linear regression model with model parameters β_0 and β_1. Another way of writing this model is

$$Y \mid x = \beta_0 + \beta_1 x + E$$

Where E[E] = 0. Our basic goal is to find estimators:

$$B_0 := \widehat{\beta}_0 = \text{ estimator for } \beta_0, \quad b_0 = \text{ estimate for } \beta_0$$
$$B_1 := \widehat{\beta}_1 = \text{ estimator for } \beta_1, \quad b_1 = \text{ estimate for } \beta_1$$

1.2 Least-Squares Estimation

Residual:

We have a random sample $(x_1,\ Y_1)$,...$(x_n,\ Y_n)$. For each measurement y_i there exists a number e_i, called the residual, such that:

$$y_i = b_0 + b_1 x_i + e_i$$

Error Sum of Squares:

$$\mathrm{SS_E} := e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

We determine the determine the estimators for β_0 and β_1 by minimizing SS_E. And the point estimates b_0 and b_1 based on this method are called least-squares estimates.

1.3 Least-Squares Estimates and Estimators

Point Estimates:

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \qquad b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$

Define:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xx} := \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} := \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$S_{xy} := \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Then we can write:

$$b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{S_{xy}}{S_{xx}}$$

Estimators:
Similar to the "maximum likelihood", replace b_0 with $\widehat{\beta_0}$ or B_0, replace $\overline{y}$ with $\overline{Y}$, ... You will get the equations for the estimators B_0 and B_1

$$B_0 = \bar{Y} - B_1\bar{x}, \quad B_1 = \frac{S_{XY}}{S_{XX}}$$

Least-Squares Estimation

Find b_0 and b_1 for the exercise data.

X	1.0	1.0	3.3	3.3	4.0	4.0	4.0	4.0	5.6	5.6	5.6	6.0	6.0	6.5	6.5
Y	1.6	1.8	1.8	1.8	2.7	2.6	2.6	2.2	3.5	2.8	2.1	3.4	3.2	3.4	3.9

2 Inferences on β_0 and β_1

2.1 Distribution of B_0, B_1 and S^2

Theorem:
Given a random sample of Y | x of size n, the following statistics follow a standard normal distribution. B_0 and B_1 are unbiased estimators, which we gain from the least squares estimation.

$$\frac{B_1 - \beta_1}{\sigma/\sqrt{\sum(x_i - \bar{x})^2}} \quad \text{and} \quad \frac{B_0 - \beta_0}{\sigma\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}}}$$

Theorem:
The variance σ^2 of Y | x is assumed to be the same for all values of x.
It turns out that the following estimator is unbiased for σ^2 and in fact follows a chi-squared distribution with n − 2 degrees of freedom.

$$\frac{(n-2)S^2}{\sigma^2} = \frac{SS_E}{\sigma^2}$$

Besides, S^2 is independent of B_0 and B_1. Analogously to the statement that the sample mean is independent of the sample variance.

2.2 Interval Estimation for β_0 and β_1

Statistic:
Hence the the following statistics follow a T -distribution with $n-2$ degrees of freedom.

$$\frac{B_1-\beta_1}{S/\sqrt{S_{xx}}} \quad \textit{and} \quad \frac{B_0-\beta_0}{S\sqrt{\sum x_k^2}/\sqrt{nS_{xx}}}$$

Confidence Intervals:
Based on the statistics, we have $100(1-\alpha)\%$ confidence intervals for β_1 and β_0:

$$B_1 \pm t_{\alpha/2,n-2}\frac{S}{\sqrt{S_{xx}}}, \quad B_0 \pm t_{\alpha/2,n-2}\frac{S\sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}}$$

Interval Estimation for β_0 and β_1

Find the 95% confidence intervals for β_0 and β_1.

X	1.0	1.0	3.3	3.3	4.0	4.0	4.0	4.0	5.6	5.6	5.6	6.0	6.0	6.5	6.5
Y	1.6	1.8	1.8	1.8	2.7	2.6	2.6	2.2	3.5	2.8	2.1	3.4	3.2	3.4	3.9

2.3 Tests for β_0 and β_1

Using the same statistics, we can also perform hypothesis tests on β_0 and β_1. Such as:

$$H_0 : \beta_0 = \beta_0^0 \quad \text{and} \quad H_0 : \beta_1 = \beta_1^0$$

An important special case is Test for Significance of Regression: We say that a regression is significant if there is statistical evidence that the slope $\beta_1 \neq 0$.

2.4 Test for Significance of Regression

Let (x_i,Y | x_i), i = 1,...,n be a random sample from Y | x.

$$H_0 : \beta_1 = 0$$

We reject at significance level α if the statistic

$$T_{n-2} = \frac{B_1}{S/\sqrt{S_{xx}}}$$

satisfies

$$|T_{n-2}| > t_{\alpha/2,n-2}$$

3 Inferences on $\mu_{Y|x}$

3.1 Distribution of $\widehat{\mu}_{Y|x}$

$$\widehat{\mu}_{Y|x} = B_0 + B_1 x = \bar{Y} - B_1\bar{x} + B_1 x = \bar{Y} + B_1(x - \bar{x})$$

So for any chosen x, it follows a normal distribution.
Besides:

$$Var[\widehat{\mu}_{Y|x}] = \frac{\sigma^2}{n} + \frac{(x-\bar{x})^2\sigma^2}{S_{xx}}$$

Hence the following statistic follows a standard-normal distribution.

$$\frac{\widehat{\mu}_{Y|x} - \mu_{Y|x}}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$$

3.2 Confidence Interval for $\mu_{Y|x}$

Statistic:
So the following statistic follows a T distribution with $n-2$ degrees of freedom.

$$\frac{\widehat{\mu}_{Y|x} - \mu_{Y|x}}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$$

Confidence Intervals:
the $100(1-\alpha)\%$ confidence interval for $\mu_{Y|x}$:

$$\widehat{\mu}_{Y|x} \pm t_{\alpha/2,n-2} S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

Confidence Interval for $\mu_{Y|x}$

1. Find the 95% confidence interval for $\mu_{Y|x}$ based on the exercise data.
2. Find the 95% confidence interval for $\mu_{Y|3.5}$

X	1.0	1.0	3.3	3.3	4.0	4.0	4.0	4.0	5.6	5.6	5.6	6.0	6.0	6.5	6.5
Y	1.6	1.8	1.8	1.8	2.7	2.6	2.6	2.2	3.5	2.8	2.1	3.4	3.2	3.4	3.9

4 Predictions on $Y|x$

1. An **estimate** is a statistical statement on the value of an unknown, but fixed, population **parameter**.
2. A **prediction** is a statistical statement on the value of an essentially **random quantity**.

Recall the general idea for we to find a confidence interval for a parameter, we can get the general idea to find a prediction interval for a random variable...

4.1 Find the Statistic

As a predictor $\widehat{Y|x}$ for the value of Y |x we use the estimator for the mean, i.e., we set

$$\widehat{Y \mid x} = \widehat{\mu}_{Y|x} = B_0 + B_1 x$$

Analyze $\widehat{\mu}_{Y|x}$ and $Y \mid x$; we know $\widehat{Y \mid x} - Y \mid x$ is normally distributed and

$$\begin{aligned} \mathrm{E}[\widehat{Y \mid x} - Y \mid x] &= \mu_{Y|x} - \mu_{Y|x} = 0 \\ \mathrm{Var}[\widehat{Y \mid x} - Y \mid x] &= \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\sigma^2 + \sigma^2 = \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\sigma^2 \end{aligned}$$

Thus, after standardizing and dividing by S/σ we obtain the T_{n2} random variable (statistic)

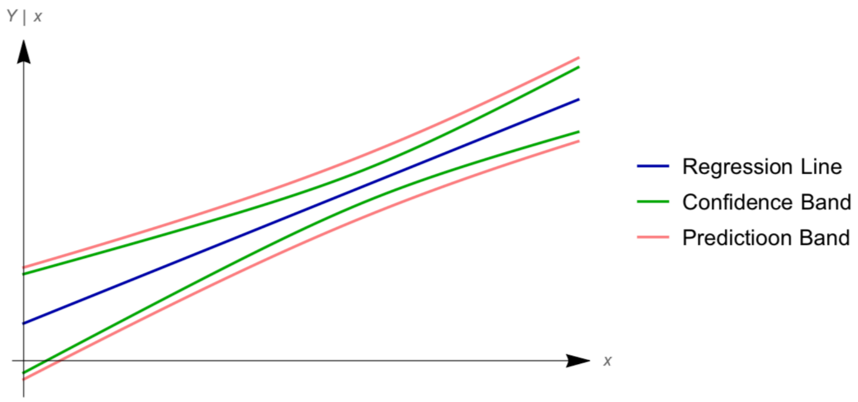
$$T_{n-2} = \frac{\widehat{Y \mid x} - Y \mid x}{S\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$$

4.2 The Prediction Interval

$100(1-\alpha)\%$ prediction interval for $Y|x$:

$$\widehat{Y \mid x} \pm t_{\alpha/2,n-2} S\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

Some comments on confidence intervals and prediction intervals...



CI and PI in a Poisson Distribution

Let X be the **total counts** in a sample of size n from a Poisson distribution with mean k, which is denoted as X $\sim$ Poisson(nk).
Let Y denote the **future total counts** that can be observed in a sample of size m from the same Poisson distribution so that Y $\sim$ Poisson(mk).
(Can be understood with "childbirth".)
Assume n is large enough.

1. Find CI for parameter k.
2. Find (one possible) PI for random variable Y.

   (The Nelson's formula: $[\lceil L \rceil, \lfloor U \rfloor]$ with $[L, U] = \widehat{Y} \pm z_{\frac{\alpha}{2}} \sqrt{m\widehat{Y}\left(\frac{1}{m} + \frac{1}{n}\right)}$)

   (Hint: Find a predictor for $\widehat{Y}$; Find a known statistic relating Y based on $\widehat{Y}$; Get the PI based on the statistic.)

HW7.2: CI and Critical Region

...

5 Model Analysis

Previously we assume our SLR model is right, then find the model parameters and get some inferences on:

1. Model parameters β_0, β_1;
2. Random variable $Y \mid x$.

Next we want to know if our linear model is appropriate.

5.1 Crucial Quantities

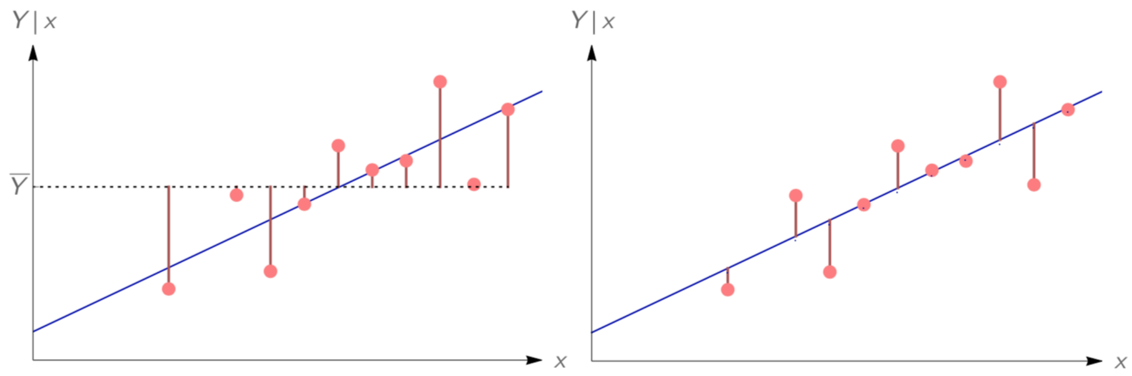
Total Sum of Squares:

$$\mathrm{SS_T} = S_{yy} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

Error Sum of Squares:

$$\mathrm{SS_E} = \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 x)\right)^2$$

$$\mathrm{SS_E} = S_{yy} - B_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$



Coefficient of Determination:

$$R^2 = \frac{SS_T - SS_E}{SS_T} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

1. R^2 expresses the proportion of the total variation in Y that is explained by the linear model.

2. R^2 is exactly the square of the estimator(22.1) for the the correlation coefficient ρ_{XY}.

   Usage1: So we can use R to **Test for Correlation Coefficient**.

3. $T_{n-2} = \frac{B_1}{\sqrt{S^2/S_{xx}}} = \frac{S_{xy}/S_{xx}}{\sqrt{SS_E/[(n-2)S_{xx}]}} = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2}$.

   The left is the statistic have used in the Test for Significance of regression.

   Usage2: So we can use R to **Test for Significance of regression**.

5.2 Test for Significance of regression

Let (X , Y) follow a bivariate normal distribution with correlation coefficient $\rho \in (-1, 1)$. Let R be the estimator(22.1) for ρ. Then

$$H_0 : \rho = 0$$

is rejected at significance level α if

$$\left| \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \right| > t_{\alpha/2,n-2}$$

Discuss on R^2

1. R^2 is large: good model because...
2. R^2 is small: means SS_E is small.

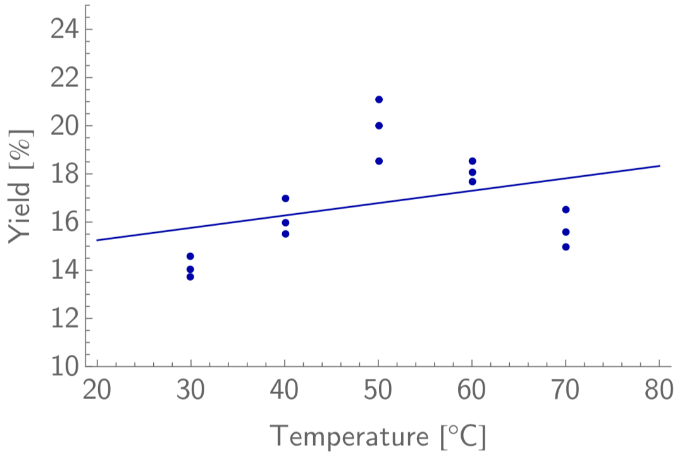
   Caused by two possible problems–

   A: due to σ^2 is very large–pure error–not model bad

   B: due to–lack-of-fit error–model bad

When R^2 is small, we test what problem it is by taking repeated measurements.

5.3 Test for Lack of Fit



Pure Error:

$$SS_{E;pe} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}Y_{ij}^2 - \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}Y_{ij}\right)^2$$

Lack of Fit Error:

$$SS_{E,lf} := SS_E - SS_{E;pe}$$

Test for Lack of Fit:

Let x_1,...,x_k be regressors and Y_{i1},Y_{i2},...,Y_{in_i}, i = 1, ... , k , the measured responses at each of the regressors. Let $SS_{E;pe}$ and $SS_{E;lf}$ be the pure error and lack-of-fit sums of squares for a linear regression model. Then

H_0 : the linear regression model is appropriate

is rejected at significance level α if the test statistic (why?)

$$F_{k-2,n-k} = \frac{SS_{E;f}/(k-2)}{SS_{E;pe}/(n-k)}$$

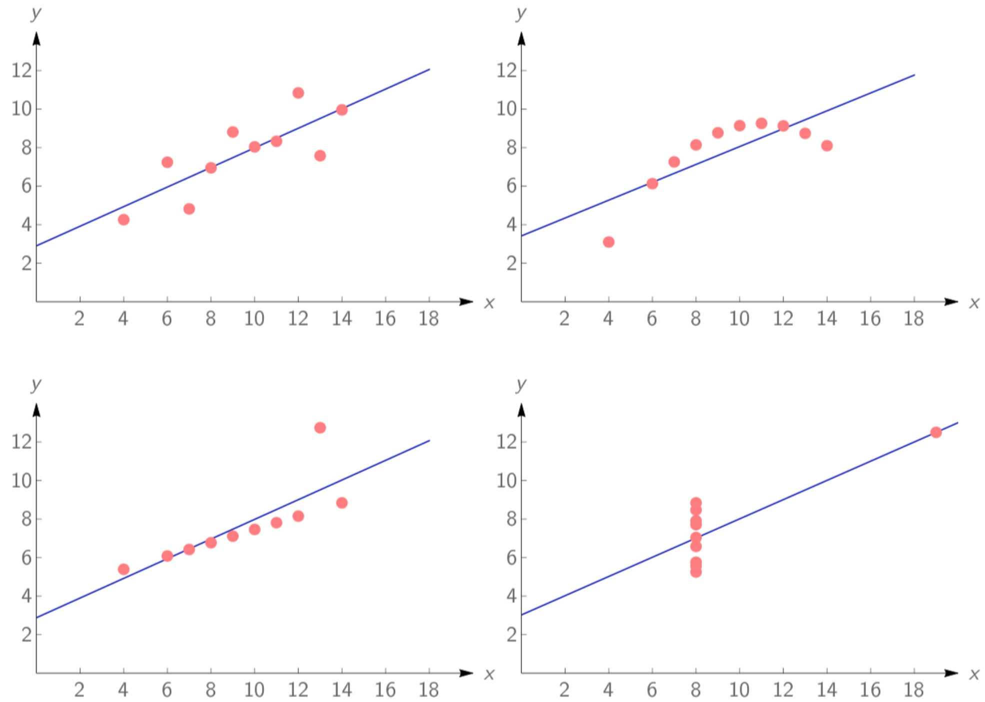
satisfies $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$.

5.4 Residual Analysis

$$e_i = Y_i - \widehat{Y}_i$$

1. Consistent with Y is of a normal distribution?
2. Consistent with Y has equal variance σ^2 for all x?
3. Does the linear model seem appropriate?

5.5 Plot the Data



Summary

Draw a graph to summary the important points you learn in SLR.

*A total Demo.